高次元小標本における平均ベクトルの推測とその周辺 (推測における統計的情報とそれに関連する話題)

AUTHOR(S):

矢田, 和善

CITATION:

矢田, 和善. 高次元小標本における平均ベクトルの推測とその周辺 (推測における統計的情報とそれに関連する話題). 数理解析研究所講究録 2011, 1758: 136-149

ISSUE DATE:

2011-08

URL:

http://hdl.handle.net/2433/171315

数理解析研究所講究録
第1758巻 2011年 136-149

高次元小標本における平均ベクトルの推測とその周辺

筑波大学・数学系　矢田　和善 (Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

1. はじめに

高次元小標本 (HDLSS) における統計的推測について，新しい方法論を提案する．HDLSS データとは，数千から数万の次元数に対して，数十から百程度の標本数からなるデータのことをいう．高次元データにおける重要な研究に，Johnstone (2001), Paul (2007) 等がある．彼らは，次元数 p と標本数 n について $n/p \to c > 0$ のもとで，標本固有値の漸近理論を研究した．それに対して，$p \to \infty$, n が固定のもとで，Hall et al. (2005), Ahn et al. (2007), Yata and Aoshima (2011a) は，HDLSS データが有する幾何学的構造を研究した．一方，Jung and Marron (2009), Yata and Aoshima (2009) は，HDLSS データに対する PCA の性質を研究した．特に，Yata and Aoshima (2009) は，従来よりも柔軟なモデル設定のもとで，PCA が一致性をもつための標本数 n の p に関するオーダー条件を導き，HDLSS データに対して PCA が不適解を起こすことを証明した．その解決策として，Yata and Aoshima (2010ab) と Yata and Aoshima (2011a) は，「クロスデータ行列法」と「ノイズ掃き出し法」という 2 つの異なるアプローチを提案し，これらの方法論による新しい PCA が，HDLSS データに対して一致性をもつ解を与えることを証明した．

本研究では，HDLSS の平均に関する各種推測に対して，予め設定される目標精度に到達するための標本数の決定方式を考える．k 個の母集団から高次元データを観測する状況を考え，母集団分布に数千単位の次元数をもつ p 次元分布を仮定する．各母集団 π_i の平均ベクトルを $\boldsymbol{\mu}_i$，共分散行列を $\boldsymbol{\Sigma}_i$ $(> \boldsymbol{O})$ とし，これらは未知であると仮定する．$\boldsymbol{\Sigma}_i$ の固有値を $\lambda_{i1} \geq \cdots \geq \lambda_{ip} > 0$ とし，適当な直交行列 $\boldsymbol{H}_i = [\boldsymbol{h}_{i1}, \cdots, \boldsymbol{h}_{ip}]$ で $\boldsymbol{\Sigma}_i = \boldsymbol{H}_i \boldsymbol{\Lambda}_i \boldsymbol{H}_i^T$, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\lambda_{i1}, \cdots, \lambda_{ip})$ と分解する．ここで，$p \to \infty$ のときにも，$\lambda_{ip} > 0$ $(i = 1, ..., k)$ を仮定する．いま，各母集団 π_i から p 次元データベクトル $\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{in_i}$ を無作為に抽出し，$\boldsymbol{z}_{ij} = (z_{i1j}, \cdots, z_{ipj})^T = \boldsymbol{\Lambda}_i^{-1/2} \boldsymbol{H}_i^T (\boldsymbol{X}_{ij} - \boldsymbol{\mu}_i)$ を定義する．ここで，$n_i = o(p)$ で，$\boldsymbol{z}_{ij}$ の成分は 4 次モーメントが一様有界と仮定する．母集団 π_i, $i = 1, ..., k$ の分布には，次の 3 つのどれか一つを仮定する：

(A-i) $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$;

(A-ii) z_{ijl}, $j = 1, ..., p$ $(l = 1, ..., n_i)$ は互いに独立である;

(A-iii) (i) $E(z_{ijl}^2 z_{isl}^2) = 1$, $E(z_{ijl} z_{isl} z_{itl} z_{iul}) = 0$, $j \neq s, t, u$, かつ (ii) $\{x_{ijl} - \mu_{ij}\}_{j \in N}$ が強定常であり ρ-mixing である．

(A-ii) は (A-i) を緩めた条件であり, (A-iii) の条件 (i) は (A-ii) を緩めた条件である. 各 $\boldsymbol{\Sigma}_i$ には以下を仮定する:

$$\textbf{(A-iv)} \quad \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^t)}{p} < \infty \ (t=1,2), \ \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i^4)}{p^2} \to 0, \ p \to \infty; \ i = 1, ..., k.$$

また, (A-iii) を仮定する場合に限り, 以下を仮定する:

$$\textbf{(A-v)} \quad \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}{p} \to c_{ij} \ (>0), \ p \to \infty; \ i, j = 1, ..., k.$$

最近, Aoshima and Yata (2011) は, 高次元小標本における重要な 8 つの推測問題を提示し, 目標精度に到達するための一連の理論と方法論を与えた. そこでは, HDLSS データが有する幾何学的構造に着目し, 高次元小標本ならではの推測理論を展開することが重要になる. さらに, Yata and Aoshima (2010c) は, 高次元小標本におけるデータ解析のために, 平均ベクトルのノルムに関する推測を考えた.

本論文では, Aoshima and Yata (2011) が与えた要求されるバンド幅をもつ信頼領域問題と要求される有意水準と検出力をもつ 2 標本問題を扱う. 各種推測には目標精度を設定し, 必要な標本数を 2 段階の標本抽出で決定することで, $p \to \infty$ における漸近的な精度を保証することを紹介する. さらに, 高次元小標本における各種推測の鍵であるパラメータ $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ の推定量をいくつか紹介し, 精度を理論的かつ数値的に比較し, 検証する. 最後に, 実際のマイクロアレイデータを用いて, 要求されるバンド幅をもつ信頼領域を構築し, その応用例も示す.

2. 要求されるバンド幅をもつ信頼領域

いま, 平均ベクトルの 1 次結合 $\boldsymbol{\mu} = \sum_{i=1}^k b_i \boldsymbol{\mu}_i$ を推定する. 各母集団から抽出される大きさ n_i の標本に基づいて $\boldsymbol{T}_{\mathrm{n}} = \sum_{i=1}^k b_i \overline{\boldsymbol{X}}_{in_i}$ を定義する. ここで, $\boldsymbol{n} = (n_1, ..., n_k)$, $\overline{\boldsymbol{X}}_{in_i} = \sum_{j=1}^{n_i} \boldsymbol{X}_{ij}/n_i$ である. 通常, 任意の $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k)$ に対して, 与えられる $d \ (>0)$ と $\alpha \in (0,1)$ による

$$P_{\boldsymbol{\theta}}(||\boldsymbol{T}_{\mathrm{n}} - \boldsymbol{\mu}|| \le d) \ge 1 - \alpha$$

という要求を満たす信頼領域を考える. この問題は Aoshima et al. (2002) によって一致解が与えられ, Aoshima and Takada (2004) によって 2 次の漸近有効性が証明された. これらは, Aoshima and Yata (2010) によって各種推測問題における 2 次漸近一致の理論に拡張され, Yata (2010) によって高次元データに対する $p \to \infty$ 漸近有効な解に拡張された. 一連の先行研究が与える解は, $n_i/p \to 0$ なる HDLSS の枠組みでは有効でないことに注意する. 正確に言えば, 要求される半径が $p \to \infty$ で有界 $(d < \infty)$ である場合, 上記要求を満たす解は存在しない. そこで, Aoshima and Yata (2011) は, 損失関数 $||\boldsymbol{T}_{\mathrm{n}} - \boldsymbol{\mu}||^2$ について, 与えられる $\delta = o(p^{1/2}) > 0$ によりバンド幅を要求した

$$R_{\boldsymbol{\Sigma}_{\mathrm{n}}} = \{\boldsymbol{\mu} \in R^p : \max\{-\delta + \boldsymbol{\Sigma}_{\mathrm{n}}, \ 0\} \le ||\boldsymbol{T}_{\mathrm{n}} - \boldsymbol{\mu}||^2 \le \delta + \boldsymbol{\Sigma}_{\mathrm{n}}\} \qquad (2.1)$$

なる領域を考えた．ここで $\mathbf{\Sigma_n} = \sum_{i=1}^{k} b_i^2 \mathrm{tr}(\mathbf{\Sigma}_i)/n_i$ である．そのとき，与えられる $\alpha \in (0,1)$ に対して，

$$P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in R_{\mathbf{\Sigma_n}}) \geq 1 - \alpha \tag{2.2}$$

なる信頼領域を求める．

2.1. 漸近正規性と標本数決定

$\mathbf{\Sigma_n} > \delta$ のとき，(2.2) 式は，中心が $\boldsymbol{T_n}$，半径が $\sqrt{\mathbf{\Sigma_n} - \delta}$ と $\sqrt{\mathbf{\Sigma_n} + \delta}$ なる 2 つの p 次元球に挟まれる領域 $R_{\mathbf{\Sigma_n}}$ において，$\boldsymbol{\mu}$ が含まれる確率を要求している．図 1 は，灰色の領域が $p = 2$ における $R_{\mathbf{\Sigma_n}}$ を表す．

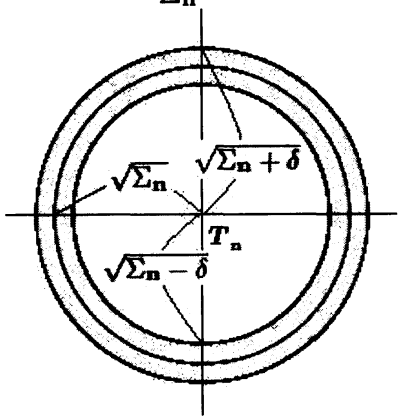

図 1. 灰色の領域: $p = 2$ における $R_{\mathbf{\Sigma_n}}$

各母集団の標本共分散行列 $\boldsymbol{S}_{in_i} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{in_i})(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_{in_i})^T$ に基づいて，$\widehat{\mathbf{\Sigma}}_{\mathbf{n}} = \sum_{i=1}^{k} b_i^2 \mathrm{tr}(\boldsymbol{S}_{in_i})/n_i$ とおく．そのとき，$||\boldsymbol{T_n} - \boldsymbol{\mu}||^2$ について Aoshima and Yata (2011) は次の定理を与えた．

定理 2.1. 母集団分布に (A-ii)，もしくは，(A-iii) かつ (A-v) を仮定する．(A-iv) と，$p \to \infty, n_i \to \infty, i = 1, ..., k$ のもとで

$$\frac{||\boldsymbol{T_n} - \boldsymbol{\mu}||^2 - \widehat{\mathbf{\Sigma}}_{\mathbf{n}}}{\sqrt{2 \sum_{i,j} b_i^2 b_j^2 \mathrm{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j)/(n_i n_j)}} \Rightarrow N(0,1)$$

が成り立つ．ここで，“$\Rightarrow$” は分布収束を意味する．

いま，$\sqrt{\sum_{i,j} b_i^2 b_j^2 \mathrm{tr}(\mathbf{\Sigma}_i \mathbf{\Sigma}_j)/(n_i n_j)} \leq \sum_{i=1}^{k} b_i^2 \sqrt{\mathrm{tr}(\mathbf{\Sigma}_i^2)}/n_i$ が成り立つことに注意する．そのとき，δ と $N(0,1)$ の上側 $\alpha/2$ 点 $z_{\alpha/2}$ によって条件を課して

$$\min \sum_{i=1}^{k} n_i \quad \text{subject to } \sqrt{2} \sum_{i=1}^{k} b_i^2 \sqrt{\mathrm{tr}(\mathbf{\Sigma}_i^2)}/n_i \leq \delta / z_{\alpha/2}$$

を求めれば，各母集団の標本数は

$$n_i \geq \frac{z_{\alpha/2}\sqrt{2}}{\delta}|b_i|\mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4}\sum_{j=1}^{k}|b_j|\mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \quad (=C_i,\ \text{say}) \tag{2.3}$$

を満たす最小の整数になる．ここで，$\delta = o(p^{1/2}) > 0$ と (A-iv) から，$C_i/p \to 0$, $p \to \infty$ が成り立ち，HDLSS の枠組みで標本数が決定されていることに注意する．このとき，Aoshima and Yata (2011) は次の定理を与えた．

定理 2.2. 母集団分布に (A-ii), もしくは，(A-iii) かつ (A-v) を仮定する．$\boldsymbol{n}$ は (2.3) 式を満たすとする．そのとき，(A-iv) と $p \to \infty$ のもとで，次が成り立つ．

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in R_{\widehat{\boldsymbol{\Sigma}}_{\mathbf{n}}}) \geq 1-\alpha.$$

2.2. 2段階推定法

(2.3) 式における $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ は未知なので，2段階推定法を考える．各 $\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ に，事前情報から得られる既知の下限 $\sigma_{i\star}$ ($\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} > \sigma_{i\star} > 0$) を仮定し，$\sigma_{i\star}/\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} \in (0,1)$, $p \to \infty$ を仮定する．いま，$\tau_\star = \min_{1\leq i\leq k}|b_i|\sqrt{\sigma_{i\star}}\sum_{j=1}^{k}|b_j|\sqrt{\sigma_{j\star}}$ とおいて，初期標本数 m を

$$m = \max\left\{4,\ \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta}\tau_\star\right]+1\right\} \tag{2.4}$$

と定義する．ここで，$[x]$ は x を越えない最大の整数を表す．各母集団から m 個の初期標本ベクトルを抽出し，$\boldsymbol{S}_{im(1)} = (m_1-1)^{-1}\sum_{j=1}^{m_1}(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{im_1})(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{im_1})^T$, $\boldsymbol{S}_{im(2)} = (m_2-1)^{-1}\sum_{j=m_1+1}^{m}(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{im_2})(\boldsymbol{X}_{ij}-\overline{\boldsymbol{X}}_{im_2})^T$ を計算する．ここで，$m_1 = [m/2]+1$, $m_2 = m-m_1$ とし，$\overline{\boldsymbol{X}}_{im_1} = \sum_{j=1}^{m_1}\boldsymbol{X}_{ij}/m_1$, $\overline{\boldsymbol{X}}_{im_2} = \sum_{j=m_1+1}^{m}\boldsymbol{X}_{ij}/m_2$ とする．いま，各母集団の標本数を

$$N_i = \max\left\{m,\ \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta}|b_i|\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^{k}|b_j|\mathrm{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\right]+1\right\} \tag{2.5}$$

で定義する．ここで，$\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})$ は $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ の不偏推定量であることに注意する (4節を参照せよ)．各母集団から追加の N_i-m 個の標本ベクトルを抽出し，初期標本と追加標本を合併して $\boldsymbol{T}_{\mathbf{N}} = \sum_{i=1}^{k}b_i\overline{\boldsymbol{X}}_{iN_i}$ と $\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = \sum_{i=1}^{k}b_i^2\mathrm{tr}(\boldsymbol{S}_{iN_i})/N_i$ を定義する．ここで，$\boldsymbol{N} = (N_1,...,N_k)$ である．このとき，Aoshima and Yata (2011) は (2.1) 式に基づいて計算される信頼領域について，次の結果を与えた．

定理 2.3. 母集団分布に (A-ii), もしくは，(A-iii) かつ (A-v) を仮定する．(A-iv) と $p \to \infty$ のもとで，次が成り立つ．

$$\liminf P_{\boldsymbol{\theta}}(\boldsymbol{\mu} \in R_{\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}}}) \geq 1-\alpha.$$

定理 2.4. 母集団分布に (A-i) を仮定する．(A-iv) と $p \to \infty$ のもとで，次が成り立つ．

$$\limsup |E_{\boldsymbol{\theta}}(N_i - C_i)| \leq 1, \quad Var_{\boldsymbol{\theta}}(N_i) = o(p^{1/2}/\delta);\ i = 1, ..., k.$$

注意 1. Yata and Aoshima (2010c) は，$||\boldsymbol{\mu}||^2$ の推定量 $\widehat{T}_{\mathrm{n}} = ||\boldsymbol{T}_{\mathrm{n}}||^2 - \widehat{\Sigma}_{\mathrm{n}}$ を用いて，p に依存する δ で区間幅を調節する

$$\boldsymbol{R}_{\mathrm{n},\delta} = \{\boldsymbol{\mu} \in R^p : \max\{\widehat{T}_{\mathrm{n}} - \delta,\ 0\} \leq ||\boldsymbol{\mu}||^2 \leq \max\{\widehat{T}_{\mathrm{n}} + \delta,\ 0\}\}$$

なる信頼領域を扱い，高次元小標本のデータ解析において使用法を示した．

2.3. シミュレーション実験

上記の 2 段階推定法 (2.4)-(2.5) によって構築した信頼領域の精度を，シミュレーション実験で検証する．$p = 1600$; $k = 2$, $b_1 = b_2 = 1$; $\delta = 5$, $\alpha = 0.05$; $m = 20$ と設定する．π_i は $N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$ と設定する．ここで，$\boldsymbol{\Sigma}_l = c_l \boldsymbol{B}(\rho_l^{|i-j|^{1/3}})\boldsymbol{B}$, $l = 1, 2$ とする．ただし，$\boldsymbol{B} = \mathrm{diag}\left(\sqrt{0.5 + 1/(p+1)}, \sqrt{0.5 + 2/(p+1)}, ..., \sqrt{0.5 + p/(p+1)}\right)$, $c_l > 0$, $\rho_l \in (0, 1)$ である．次の 3 つの場合について，2000 回のシミュレーション結果を纏めたものが表 1 である．(i) $(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.3)$; (ii) $(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.4)$; (iii) $(c_1, c_2) = (1, 1.5)$, $(\rho_1, \rho_2) = (0.3, 0.3)$. ここでは割愛するが，設定を変えて実験をしたときも，2 段階推定法による信頼領域が数千の次元数で要求精度を満たすことを確認した．

表 1. 2 段階推定法で構築した信頼領域の精度 ($p = 1600$; $\delta = 5$, $\alpha = 0.05$)

		$\overline{N}$	$\overline{N} - C$	$Var(N)$	$\overline{P}$	$s(\overline{P})$
		$(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.3)$				
C	116.29	117.00	0.72	47.81	0.943	0.00518
C_1	58.14	58.50	0.36	15.13		
C_2	58.14	58.50	0.36	14.83		
		$(c_1, c_2) = (1, 1)$, $(\rho_1, \rho_2) = (0.3, 0.4)$				
C	131.66	132.24	0.58	69.54	0.950	0.00487
C_1	61.87	62.17	0.30	16.60		
C_2	69.79	70.07	0.28	27.08		
		$(c_1, c_2) = (1, 1.5)$, $(\rho_1, \rho_2) = (0.3, 0.3)$				
C	143.89	144.21	0.32	74.89	0.946	0.00505
C_1	64.68	64.88	0.20	17.53		
C_2	79.21	79.33	0.12	29.48		

3. 要求される有意水準と検出力をもつ2標本問題

2つの母集団の平均ベクトル $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ について，次の検定を考える：

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{3.1}$$

Bai and Saranadasa (1996), Chen and Qin (2010) は，$p \to \infty$ のときに (3.2) で与えられる検定統計量を提案した．Aoshima and Yata (2011) は，$n_i/p \to 0$ において要求される有意水準と検出力をもつ検定方式を与えた．いま，$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とおく．与えられる α, $\beta \in (0, 1/2)$, $\Delta_L = o(p^{1/2})$ (> 0) に対して，有意水準 (size)$\leq \alpha$, $\Delta \geq \Delta_L$ のときの検出力 (power)$\geq 1 - \beta$ となるような検定方式を求める．

3.1. 漸近正規性と標本数決定

各母集団から抽出される大きさ n_i の標本に基づいて，Δ の推定量

$$\widetilde{T}_{\mathbf{n}} = \sum_{i=1}^{2} \frac{\sum_{j \neq j'}^{n_i} \boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij'}}{n_i(n_i - 1)} - 2\frac{\sum_{j=1}^{n_1} \sum_{j'=1}^{n_2} \boldsymbol{X}_{1j}^T \boldsymbol{X}_{2j'}}{n_1 n_2} \tag{3.2}$$

を考える．ここで，$E_{\boldsymbol{\theta}}(\widetilde{T}_{\mathbf{n}}) = \Delta$,

$$Var_{\boldsymbol{\theta}}(\widetilde{T}_{\mathbf{n}}) = \sum_{i=1}^{2} \frac{2}{n_i(n_i - 1)} \mathrm{tr}(\boldsymbol{\Sigma}_i^2) + \frac{4}{n_1 n_2} \mathrm{tr}(\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2) + \sum_{i=1}^{2} \frac{4}{n_i} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

である．このとき，Aoshima and Yata (2011) は次の結果を与えた．

定理 3.1. 母集団分布に (A-ii), もしくは，(A-iii) かつ (A-v) を仮定する．さらに，$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/n_i)$, $i = 1, 2$ も仮定する．そのとき，(A-iv) のもと $p \to \infty$, $n_i \to \infty$, $i = 1, 2$ のとき，次が成り立つ．

$$\frac{\widetilde{T}_{\mathbf{n}} - \Delta}{\sqrt{Var_{\boldsymbol{\theta}}(\widetilde{T}_{\mathbf{n}})}} \Rightarrow N(0, 1).$$

注意 2. Chen and Qin (2010) は，異なる条件のもとで $\widetilde{T}_{\mathbf{n}}$ の漸近正規性を示した．

各母集団の標本数を

$$n_i \geq \frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L} \mathrm{tr}(\boldsymbol{\Sigma}_i^2)^{1/4} \sum_{j=1}^{2} \mathrm{tr}(\boldsymbol{\Sigma}_j^2)^{1/4} \quad (= C_i, \text{ say}) \tag{3.3}$$

を満たす整数とし，(3.2) 式の検定統計量に基づく検定方式を

$$H_0 \text{ を棄却} \iff \widetilde{T}_{\mathbf{n}} > \frac{\Delta_L z_\alpha}{z_\alpha + z_\beta} \tag{3.4}$$

で定義する．このとき, Aoshima and Yata (2011) は次の定理を与えた．

定理 3.2. 母集団分布に (A-ii), もしくは，(A-iii) かつ (A-v) を仮定する．n_1, n_2 は (3.3) を満たすものとする．(A-iv) と $p \to \infty$ のもとで，検定方式 (3.4) は次が成り立つ．

$$\limsup \text{size} \leq \alpha \quad \text{and} \quad \liminf \text{power}(\Delta_L) \geq 1 - \beta. \tag{3.5}$$

ただし，$\text{power}(\Delta_L)$ は $\Delta = \Delta_L$ における検出力である．

3.2. 2段階推定法

(3.3) 式における $\text{tr}(\Sigma_i^2)$ は未知なので，2段階推定法を考える．各 $\sqrt{\text{tr}(\Sigma_i^2)}$ に，事前情報から得られる既知の下限 $\sigma_{i\star}$ ($\sqrt{\text{tr}(\Sigma_i^2)} > \sigma_{i\star} > 0$) を仮定し，$\sigma_{i\star}/\sqrt{\text{tr}(\Sigma_i^2)} \in (0,1)$, $p \to \infty$ を仮定する．いま，$\tau_\star = \min_{1 \leq i \leq 2} \sqrt{\sigma_{i\star}} \sum_{j=1}^2 \sqrt{\sigma_{j\star}}$ とおいて，初期標本数 m を

$$m = \max\left\{4, \left[\frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L}\tau_\star\right] + 1\right\} \tag{3.6}$$

と定義する．各母集団から m 個の初期標本ベクトルを抽出し，2節と同様に $\boldsymbol{S}_{im(1)}$, $\boldsymbol{S}_{im(2)}$ を計算する．いま，各母集団の標本数を

$$N_i = \max\left\{m, \left[\frac{(z_\alpha + z_\beta)\sqrt{2}}{\Delta_L}\text{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})^{1/4}\sum_{j=1}^2 \text{tr}(\boldsymbol{S}_{jm(1)}\boldsymbol{S}_{jm(2)})^{1/4}\right] + 1\right\} \tag{3.7}$$

で定義する．各母集団から追加の $N_i - m$ 個の標本ベクトルを抽出し，初期標本と追加標本を合併して $\widetilde{T}_{\mathbf{N}}$ を計算する．そのとき，検定方式を

$$H_0 \text{ を棄却} \iff \widetilde{T}_{\mathbf{N}} > \frac{\Delta_L z_\alpha}{z_\alpha + z_\beta} \tag{3.8}$$

で定義する．このとき, Aoshima and Yata (2011) は次の定理を与えた．

定理 3.3. 母集団分布に (A-ii), もしくは，(A-iii) かつ (A-v) を仮定する．このとき，(A-iv) と $p \to \infty$ のもと，検定方式 (3.8) に (3.5) が成り立つ．

定理 3.4. 母集団分布に (A-i) を仮定する．(A-iv) と $p \to \infty$ のもとで，次が成り立つ．

$$\limsup |E_{\boldsymbol{\theta}}(N_i - C_i)| \leq 1, \quad Var_{\boldsymbol{\theta}}(N_i) = o(p^{1/2}/\Delta_L);\ i = 1, 2.$$

4. 高次元小標本における $\text{tr}(\Sigma_i^2)$ の推定量

本節では，高次元小標本の精度保証を有する推測において，しばしば重要とな

る $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ の推定を考える．以降，母集団を表す添え字 i は省く．単純な推定量 $\mathrm{tr}(\boldsymbol{S}_n^2)$ は，高次元において非常に大きなバイアスを生じることに注意する．いま，$n_{(1)}=[n/2]+1$, $n_{(2)}=n-n_{(1)}$ とし，

$$\boldsymbol{S}_{n(1)}=(n_{(1)}-1)^{-1}\sum_{j=1}^{n_{(1)}}(\boldsymbol{X}_j-\overline{\boldsymbol{X}}_{n_{(1)}})(\boldsymbol{X}_j-\overline{\boldsymbol{X}}_{n_{(1)}})^T,$$

$$\boldsymbol{S}_{n(2)}=(n_{(2)}-1)^{-1}\sum_{j=n_{(1)}+1}^{n}(\boldsymbol{X}_j-\overline{\boldsymbol{X}}_{n_{(2)}})(\boldsymbol{X}_j-\overline{\boldsymbol{X}}_{n_{(2)}})^T$$

とおく．ただし，$\overline{\boldsymbol{X}}_{n_{(1)}}=\sum_{j=1}^{n_{(1)}}\boldsymbol{X}_j/n_{(1)}$, $\overline{\boldsymbol{X}}_{n_{(2)}}=\sum_{j=n_{(1)}+1}^{n}\boldsymbol{X}_j/n_{(2)}$ である．そのとき，Yata (2010) は，$E_{\boldsymbol{\theta}}\{\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})\}=\mathrm{tr}(\boldsymbol{\Sigma}^2)$ なる不偏推定量 $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$ を与えた．いま，$Var_{\boldsymbol{\theta}}(z_{jl}^2)=M_j(<\infty)$, $j=1,...,p$ とおく．ただし，(A-i) のもとで，$M_j=2$, $j=1,...,p$ であることに注意する．このとき，母集団分布が (A-iii) の条件 (i) のもとで，$p\to\infty$, $n\to\infty$ のとき

$$Var_{\boldsymbol{\theta}}\left(\frac{\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right)=\frac{8}{n^2}(1+o(1))+\frac{4\sum_{j=1}^{p}\lambda_j^4M_j(1+o(1))}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2n} \tag{4.1}$$

が主張でき，$n/p\to 0$ なる高次元小標本の枠組みで一致性が主張できる．一方，母集団分布が (A-iii) の条件 (i) が仮定できないもとで，$p\to\infty$, $n\to\infty$ のとき

$$Var_{\boldsymbol{\theta}}\left(\frac{\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right)=O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma})^4}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2n^2}\Big)+O(n^{-1}) \tag{4.2}$$

と評価できる．

Bai and Saranadasa (1996), Srivastava (2005) は，推定量 $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})=c_n^{-1}\{\mathrm{tr}(\boldsymbol{S}_n^2)-\mathrm{tr}(\boldsymbol{S}_n)^2/(n-1)\}$ を与えた．ここで，$c_n=(n-2)(n+1)/(n-1)^2$ である．そのとき，母集団分布に (A-i) を仮定できれば，$E_{\boldsymbol{\theta}}\{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})\}=\mathrm{tr}(\boldsymbol{\Sigma}^2)$ となり，$p\to\infty$, $n\to\infty$ のとき

$$Var_{\boldsymbol{\theta}}\left(\frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right)=\frac{4}{n^2}(1+o(1))+\frac{8\mathrm{tr}(\boldsymbol{\Sigma}^4)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2n}(1+o(1)) \tag{4.3}$$

が主張できる．それゆえ，(4.1) と (4.3) より，(A-i) のもとでは $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})$ が $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$ に比べ，漸近的に分散が小さい．しかしながら，母集団分布に (A-i) が仮定できない非正規の場合には，推定量 $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})$ の不偏性は主張できず，高次元のもとで非常に大きなバイアスを生じる．さらには，$\boldsymbol{z}_j$ の成分について8次モーメントの一様有界性が仮定できない場合，$Var_{\boldsymbol{\theta}}(\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})/\mathrm{tr}(\boldsymbol{\Sigma}^2))<\infty$ さえ保証しない．Yata (2010) が提案した推定量 $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$ は，$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}_n^2})$ に比べて，非正規のもとで非常に頑健な推定量であるといえる．

一方, Chen and Qin (2010) は次のような $\mathrm{tr}(\Sigma^2)$ の推定量を与えた.

$$\mathrm{tr}(\widehat{\Sigma^2_{CQ}}) = (n(n-1))^{-1}\mathrm{tr}\{\sum_{j\neq k}^{n}(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_{n(j,k)})\boldsymbol{X}_j^T(\boldsymbol{X}_k - \overline{\boldsymbol{X}}_{n(j,k)})\boldsymbol{X}_k^T\}.$$

ただし, $\overline{\boldsymbol{X}}_{n(j,k)}$ は $\boldsymbol{X}_j$ と $\boldsymbol{X}_k$ を除いた $n-2$ 個のデータにおける標本平均である. しかしながら, $E_{\boldsymbol{\theta}}\{\mathrm{tr}(\widehat{\Sigma^2_{CQ}})\} = \mathrm{tr}(\Sigma^2) + \boldsymbol{\mu}^T\Sigma\boldsymbol{\mu}/(n-2)$ となり, $||\boldsymbol{\mu}||^2 = O(p)$ など $||\boldsymbol{\mu}||^2$ が大きい場合は非常に大きなバイアスを生じることに注意する.

4.1. 新しい $\mathrm{tr}(\Sigma^2)$ の推定量における漸近的性質

本節では, Yata and Aoshima (2011b) が提案した新しい $\mathrm{tr}(\Sigma^2)$ の推定量における漸近的性質を考える. いま, $k = 3, ..., 2n-1$ において

$$\boldsymbol{V}_{nk(1)} = \begin{cases} \{[k/2] - n_{(1)} + 1, ..., [k/2]\} & \text{if } [k/2] \geq n_{(1)}, \\ \{1, ..., [k/2]\} \cup \{n_{(2)} + [k/2] + 1, ..., n\} & \text{otherwise,} \end{cases}$$

$$\boldsymbol{V}_{nk(2)} = \begin{cases} \{[k/2] + 1, ..., [k/2] + n_{(2)}\} & \text{if } [k/2] \leq n_{(1)}, \\ \{1, ..., [k/2] - n_{(1)}\} \cup \{[k/2] + 1, ..., n\} & \text{otherwise} \end{cases}$$

なる集合 $\boldsymbol{V}_{nk(1)}$, $\boldsymbol{V}_{nk(2)}$ を定義する. ここで, $|\boldsymbol{S}|$ は集合 $\boldsymbol{S}$ の成分の数を表すとし, $|\boldsymbol{V}_{nk(l)}| = n_{(l)}$, $l = 1, 2$, $\boldsymbol{V}_{nk(1)} \cap \boldsymbol{V}_{nk(2)} = \boldsymbol{\emptyset}$, $\boldsymbol{V}_{nk(1)} \cup \boldsymbol{V}_{nk(2)} = \{1, ..., n\}$ に注意する. さらに,

$$\overline{\boldsymbol{X}}_{nk(1)} = \sum_{l\in\boldsymbol{V}_{nk(1)}} \boldsymbol{x}_l/n_{(1)}, \qquad \overline{\boldsymbol{X}}_{nk(2)} = \sum_{l\in\boldsymbol{V}_{nk(2)}} \boldsymbol{x}_l/n_{(2)}$$

とおく. そのとき, Yata and Aoshima (2011b) は $\mathrm{tr}(\Sigma^2)$ の推定量

$$\begin{aligned} \mathrm{tr}(\widetilde{\Sigma^2_n}) &= 2u_n \sum_{j<j'}^{n} \frac{\left((\boldsymbol{X}_j - \overline{\boldsymbol{X}}_{nj+j'(1)})^T(\boldsymbol{X}_{j'} - \overline{\boldsymbol{X}}_{nj+j'(2)})\right)^2}{n(n-1)} \\ &= 2u_n \sum_{k=3}^{2n-1} \sum_{j=1}^{[k/2]} \frac{\left((\boldsymbol{X}_j - \overline{\boldsymbol{X}}_{nk(1)})^T(\boldsymbol{X}_{k-j} - \overline{\boldsymbol{X}}_{nk(2)})\right)^2}{n(n-1)} \end{aligned} \tag{4.4}$$

を与えた. ここで, $u_n = n_{(1)}n_{(2)}/((n_{(1)}-1)(n_{(2)}-1))$ である. このとき, $j < j'$ において, $(\boldsymbol{X}_j - \overline{\boldsymbol{X}}_{nj+j'(1)})$ と $(\boldsymbol{X}_{j'} - \overline{\boldsymbol{X}}_{nj+j'(2)})$ は独立であり, $E_{\boldsymbol{\theta}}\{\mathrm{tr}(\widetilde{\Sigma^2_n})\} = \mathrm{tr}(\Sigma^2)$ が常に主張できる. 分散について, 母集団分布が (A-iii) の条件 (i) のもとで, $p \to \infty$, $n \to \infty$ のとき

$$Var_{\boldsymbol{\theta}}\left(\frac{\mathrm{tr}(\widetilde{\Sigma^2_n})}{\mathrm{tr}(\Sigma^2)}\right) = \frac{4}{n^2}(1 + o(1)) + \frac{4\sum_{j=1}^{p}\lambda_j^4 M_j}{\mathrm{tr}(\Sigma^2)^2 n}(1 + o(1)) \tag{4.5}$$

が主張できる．それゆえ，(4.1), (4.3) と (4.5) より，$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)$ は $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$ に比べ，漸近的に分散が小さく，(A-i) のもとでも $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n^2)$ と同等な漸近分散をもつ．よって，$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)$ は頑健かつ漸近的に分散が小さい $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の不偏推定量といえる．一方で，母集団分布が (A-iii) の条件 (i) が仮定できないもとで，$p \to \infty$, $n \to \infty$ のとき

$$Var_\theta\left(\frac{\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)}{\mathrm{tr}(\boldsymbol{\Sigma}^2)}\right) = O\Big(\frac{\mathrm{tr}(\boldsymbol{\Sigma})^4}{\mathrm{tr}(\boldsymbol{\Sigma}^2)^2 n^2}\Big) + O(n^{-1}) \tag{4.6}$$

も主張できる．

注意 3. (2.5) もしくは (3.7) において，$\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})$ の代わりに (4.4) に基づく推定量 $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_{im}^2)$ を用いても，定理 2.3-2.4 もしくは定理 3.3-3.4 が成り立つ．

4.2. シミュレーション実験

3 つの $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の推定量，$\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})$, $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n^2)$ と $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)$ の精度を，シミュレーション実験で検証する．

まず $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n^2)$ が不偏推定量となる $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ のもとでデータを発生させる．$n = 50$, $p = 600(200)1600$, $\boldsymbol{\Sigma} = (0.3^{|i-j|^{1/3}})$ と設定する．図 2.1 は，A : $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})/\mathrm{tr}(\boldsymbol{\Sigma}^2)$, B : $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n^2)/\mathrm{tr}(\boldsymbol{\Sigma}^2)$, C : $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)/\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の値について，それぞれ 1000 回のシミュレーション実験を行い，その平均値をプロットしたものであり，図 2.2 は，A,B,C の不偏分散の値をプロットしたものである．

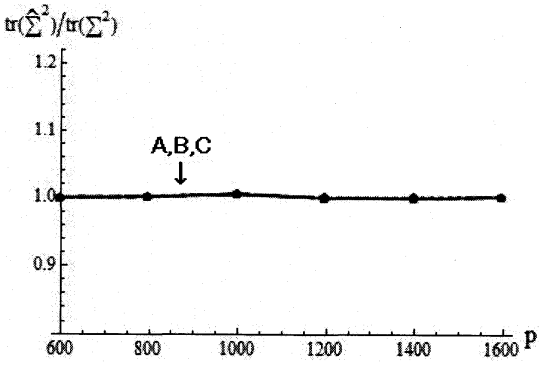

図 2.1. A,B,C の平均値.

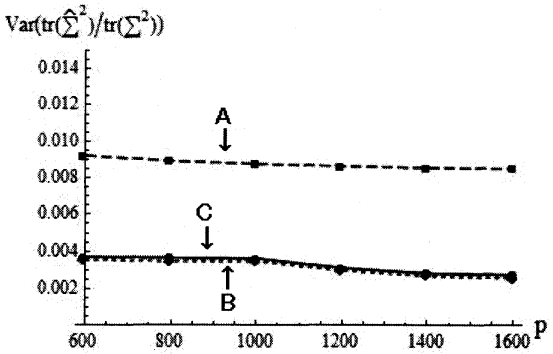

図 2.2. A,B,C の不偏分散.

図 2.1 から分かるように，(A-i) のもと $\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の推定について，すべての推定量とも不偏性を持つことが確認できた．しかしながら，図 2.2 より，推定量の分散は A が B,C に比べ多少大きくなる．一方，(4.3) と (4.5) より，B と C は理論的に同等な分散をもち，それが数値的にも確認できた．

次にデータが非正規分布に従うもとで，シミュレーション実験を行う．平均 $\mathbf{0}$, 共分散行列 $\boldsymbol{\Sigma} = (0.3^{|i-j|^{1/3}})$, 自由度 $\nu = 20(20)120$ の $p = 1000$ 次元の t 分布の乱数を生成する．図 3.1 は，A : $\mathrm{tr}(\boldsymbol{S}_{n(1)}\boldsymbol{S}_{n(2)})/\mathrm{tr}(\boldsymbol{\Sigma}^2)$, B : $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_n^2)/\mathrm{tr}(\boldsymbol{\Sigma}^2)$, C :

$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)/\mathrm{tr}(\boldsymbol{\Sigma}^2)$ の値について，それぞれ1000回のシミュレーション実験を行い，その平均値をプロットしたものであり，図3.2は，A,B,Cの不偏分散の値をプロットしたものである．

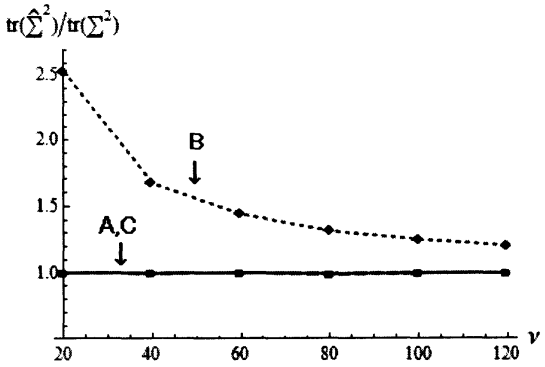

図3.1. A,B,Cの平均値.

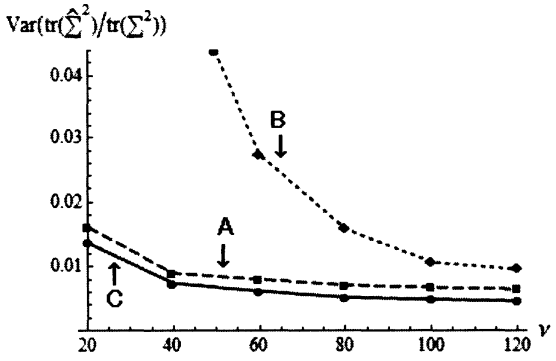

図3.2. A,B,Cの不偏分散.

いま，自由度νが大きくなるにつれ，t分布は正規分布に近づくことに注意をすると，これらの図から分かるように，自由度が小さく，正規分布から離れた分布においては，Bは不偏性を持たず，大きなバイアスが生じることが確認できる．νが大きく，正規分布に近い分布においては，Bの不偏性は多少回復する．一方で，どんなνの値においてもAとCは不偏性を持つことが確認でき，Cの方が分散が小さいことも確認できる．

これらの結果からも$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_n^2)$が不偏であり，頑健かつ漸近的に分散が小さい$\mathrm{tr}(\boldsymbol{\Sigma}^2)$の推定量といえる．

5. マイクロアレイデータ解析

本節では，Chiaretti et al. (2004) の12625($=p$)遺伝子からなるマイクロアレイデータを用いて，2節で紹介した要求されるバンド幅をもつ信頼領域の解析例を与える．このマイクロアレイデータはπ_1: B-cell と π_2: T-cellの2タイプの腫瘍のデータからなる．

$\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ $(b_1 = 1,\ b_2 = -1)$, $\alpha = 0.05$, $\delta = 100$ と設定する．B-cellにおいて，$\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)} > 300$, T-cellにおいて，$\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_2^2)} > 300$ と仮定する．よって，$\sigma_{1\star} = 300$, $\sigma_{2\star} = 300$ と設定し，$\tau_\star = \min_{i=1,2} \sigma_{i\star}^{1/2}(\sigma_{1\star}^{1/2} + \sigma_{2\star}^{1/2}) = 600$ を得る．そのとき，(2.4)より，初期標本数mを

$$m = \left\{4,\ \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta}\tau_\star\right] + 1\right\} = 17$$

とする．よって，各母集団から$m(=17)$個の初期標本ベクトルを抽出し，(4.4)に基づき，$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_{1m}^2) = 578^2$, $\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}}_{2m}^2) = 423^2$ を得る．

いま，(2.5)において$\mathrm{tr}(\boldsymbol{S}_{im(1)}\boldsymbol{S}_{im(2)})$の代わりに，より分散が小さい不偏推定量

$\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}_{im}^2})$ を用いて, 各母集団の標本数を

$$N_1 = \max\left\{m,\ \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta}\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}_{1m}^2})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}_{jm}^2})^{1/4}\right]+1\right\} = 30,$$

$$N_2 = \max\left\{m,\ \left[\frac{z_{\alpha/2}\sqrt{2}}{\delta}\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}_{2m}^2})^{1/4}\sum_{j=1}^{2}\mathrm{tr}(\widetilde{\boldsymbol{\Sigma}_{jm}^2})^{1/4}\right]+1\right\} = 26$$

と決定する. よって, B-cellから13個の追加標本, T-cellから9個の追加標本をそれぞれ抽出し, 初期標本と追加標本を合併して $\boldsymbol{T}_{\mathbf{N}} = \overline{\boldsymbol{X}}_{1N_1} - \overline{\boldsymbol{X}}_{2N_2} = (-0.120, -0.012, 0.033, ..., 0.102, 0.060, 0.160)^T$ と $\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}} = \sum_{i=1}^{2}\mathrm{tr}(\boldsymbol{S}_{iN_i})/N_i = 175.2$ を得る. そのとき, 漸近的に

$$P_{\boldsymbol{\theta}}\left(\boldsymbol{\mu} \in R_{\widehat{\boldsymbol{\Sigma}}_{\mathbf{N}}}\right) = P_{\boldsymbol{\theta}}(75.2 \leq ||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{\mu}||^2 \leq 275.2) \geq 0.95 \tag{5.1}$$

が保証される.

ここで, 信頼領域 (5.1) を用いた応用例を与える. まず信頼領域 (5.1) を用いて, $\boldsymbol{\mu} = \mathbf{0}$ $(\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2)$ が成り立つか確認する. いま, $||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{\mu}||^2 = ||\boldsymbol{T}_{\mathbf{N}}||^2 = 1744$ より, 信頼領域 (5.1) に $\boldsymbol{\mu} = \mathbf{0}$ は含まれない. よって, $\boldsymbol{\mu} \neq \mathbf{0}$ が保証される.

次に, $\boldsymbol{T}_{\mathbf{N}} = (T_{1\mathbf{N}}, ..., T_{p\mathbf{N}})^T$ とし, $\gamma > 0$ とおき, 変数選択手法

$$T_{j\mathbf{N}(*)} = \begin{cases} T_{j\mathbf{N}} & \text{if } |T_{j\mathbf{N}}| \geq \gamma, \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

を考える. まず $\gamma = 0.4$ の場合を考える. そのとき, $\boldsymbol{T}_{\mathbf{N}(*)} = (T_{1\mathbf{N}(*)}, ..., T_{p\mathbf{N}(*)})^T$ とし,

$$\boldsymbol{T}_{\mathbf{N}(*)} = (0, 0, 0, 0, 0, 0.566, ..., 0, 0, 0)^T$$

を得る. ここで, $\boldsymbol{T}_{\mathbf{N}(*)}$ の 0 以外の成分は 1795 個であり, 12625 変数 (遺伝子) から 1795 変数 (遺伝子) に削減できた. このとき, $\boldsymbol{\mu} = \boldsymbol{T}_{\mathbf{N}(*)}$ を確認する. ここで, $||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{\mu}||^2 = ||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{T}_{\mathbf{N}(*)}||^2 = 271.9$ より, 信頼領域 (5.1) に $\boldsymbol{\mu} = \boldsymbol{T}_{\mathbf{N}(*)}$ は含まれる. よって, $\boldsymbol{T}_{\mathbf{N}(*)}$ を $\boldsymbol{\mu}$ の一つの推定量と考えることができ, $\gamma = 0.4$ における変数選択手法 (5.2) が変数 (遺伝子) を有効に選択できていると考えられる.

次に $\gamma = 0.8$ の場合を考えてみる. そのとき, $\boldsymbol{T}_{\mathbf{N}(*)}$ を求め, $\boldsymbol{T}_{\mathbf{N}(*)}$ の 0 以外の成分は 533 個となり, 533 変数 (遺伝子) に削減できた. このとき, $\boldsymbol{\mu} = \boldsymbol{T}_{\mathbf{N}(*)}$ を確認すると, $||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{\mu}||^2 = ||\boldsymbol{T}_{\mathbf{N}} - \boldsymbol{T}_{\mathbf{N}(*)}||^2 = 663.0$ より, 信頼領域 (5.1) に $\boldsymbol{\mu} = \boldsymbol{T}_{\mathbf{N}(*)}$ は含まれない. よって, $\gamma = 0.8$ における変数選択手法 (5.2) は変数 (遺伝子) を有効に選択出来ていないと考えられる. それは, $\gamma = 0.4$ の場合に比べ, 遺伝子を大幅に削減しており, それゆえ, 多くの選択されるべき重要な遺伝子を削除していることが原因として考えられる.

謝辞　本研究は，科学研究費補助金 基盤研究 (B) 22300094 研究代表者: 青嶋 誠「高次元データの理論と方法論の総合的研究」から，研究助成を受けています.

REFERENCES

Ahn, J., Marron, J. S., Muller, K. M. and Chi, Y.-Y. (2007). The High-Dimension, Low-Sample-Size Geometric Representation Holds Under Mild Conditions. *Biometrika* 94: 760–766.

Aoshima, M. and Takada, Y. (2004). Asymptotic Second-Order Efficiency for Multivariate Two-Stage Estimation of a Linear Function of Normal Mean Vectors. *Seq. Anal.* 23(3): 333-353.

Aoshima, M., Takada, Y. and Srivastava, M. S. (2002). A Two-Stage Procedure for Estimating a Linear Function of k Multinormal Mean Vectors When Covariance Matrices and Unknown. *J. Statist. Plann. Inference* 100: 109-119.

Aoshima, M. and Yata, K. (2010). Asymptotic Second-Order Consistency for Two-Stage Estimation Methodologies and Its Applications. *Ann. Inst. Statist. Math.* 62: 571-600.

Aoshima, M. and Yata, K. (2011). Two-Stage Procedures for High-Dimensional Data. *Seq. Anal.*, to appear (*Editor's Special Invited Article*).

Bai, Z. and Sarandasa, H. (1996). Effect of High Dimension: By an Example of a Two Sample Problem. *Statist. Sin.* 6: 311–329.

Chen, S. X. and Qin, Y.-L. (2010). A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *Ann. Statist.* 38: 808–835.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene Expression Profile of Adult T-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival, *Blood* 103: 2771-2778.

Hall, P., Marron, J. S. and Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data. *J. Roy. Statist. Soc. B* 67: 427-444.

Johnstone, I. M. (2001). On the Distribution of the Largest Eigenvalue in Principal Components Analysis. *Ann. Statist.* 29: 295-327.

Jung, S. and Marron, J. S. (2009). PCA Consistency in High Dimension, Low Sample Size Context. *Ann. Statist.* 37: 4104-4130.

Paul, D. (2007). Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. *Statist. Sin.* 17: 1617-1642.

Srivastava, M. S. (2005). Some Tests Concerning the Covariance Matrix in High-Dimensional Data. *J. Japan Statist. Soc.* 35: 251-272.

Yata, K. (2010). Effective Two-Stage Estimation for a Linear Function of High-Dimensional Gaussian Means. *Seq. Anal.* 29: 463-482.

Yata, K. and Aoshima, M. (2009). PCA Consistency for Non-Gaussian Data in High Dimension, Low Sample Size Context. *Commun. Statist.-Theory Meth., Special Issue Honoring S. Zacks* (ed. N. Mukhopadhyay) 38: 2634-2652.

Yata, K. and Aoshima, M. (2010a). Effective PCA for High-Dimension, Low-Sample-Size Data with Singular Value Decomposition of Cross Data Matrix. *J. Multivariate Anal.* 101: 2060-2077.

Yata, K. and Aoshima, M. (2010b). Intrinsic Dimensionality Estimation of High Dimension, Low Sample Size Data with *d*-asymptotics. *Commun. Statist.-Theory Meth., Special Issue Honoring M. Akahira* (ed. M. Aoshima) 39: 1511-1521.

Yata, K. and Aoshima, M. (2010c). Inference on High-Dimensional Mean Vectors with Fewer Observations Than the Dimension, submitted.

Yata, K. and Aoshima, M. (2011a). Effective PCA for High-Dimension, Low-Sample-Size Data with Noise Reduction via Geometric Representations. *J. Multivariate Anal.*, revised.

Yata, K. and Aoshima, M. (2011b). Correlation Tests for High-Dimensional Data, submitted.